Proyecto de Innovación Docente. Universidad de Salamanca, 2015

Curso Virtual sobre Técnicas Análisis de Redes Sociales, Fundamentos y Aplicación

Memoria de actuaciones.

Carlos G. Figuerola
José Luis Alonso Berrocal
Angel Zazo Rodríguez

1.- Introducción

Las *Técnicas de Análisis de Redes Sociales* (SNA), a pesar de su nombre, son una aplicación generalista de la Teoría de Grafos no circunscrita necesariamente ni a las redes o vínculos que establecen las personas entre sí, ni tampoco a fenómenos cono *Twitter, Facebook*, y similares; aunque también pueden aplicarse en estos campos.

Debido a ese carácter generalista, se aplican en multitud de campos y debido a eso trabajan con tales técnicas o muestran interés por ellas investigadores de áreas muy diversas. Esta diversidad plantea varias cuestiones interesantes desde el punto de vista del presente proyecto:

- los orígenes académicos de quienes se interesan por SNA son muy variados, y las capacidades para entender y aplicar tales modelos y técnicas es muy diversa
- más allá de las bases teóricas de SNA, sus posibilidades de aplicación y el significado real de índices, coeficientes, medidas, etc. que SNA nos permite obtener son difíciles de apreciar y comprender sin una experiencia más o menos dilatada en la aplicación de estos modelos y técnicas.

De ahí la necesidad de contar con ejemplos procedentes de casos reales que no sólo permitan aplicar estas técnicas, sino que, además, la utilización de técnicas SNA produzca conocimiento nuevo y permite obtener conclusiones útiles e interesantes.

Existen, en la actualidad, numerosas herramientas informáticas que permiten modelar datos aplicando SNA y aplicar las mencionadas técnicas. Sin embargo, para utilizar tales herramientas informáticas, y para enseñar a utilizarlas, es preciso disponer de colecciones de datos interesantes, que permitan apreciar la potencia de tales ténicas e instrumentos informáticos y, por supuesto, que permitan llevar a cabo un aprendizaje eficaz.

El objetivo principal de este Proyecto ha sido producir una serie de colecciones de datos para ser utilizados en la enseñanza de las técnicas SNA. Estas colecciones de datos debían diseñarse atendiendo a varios requerimientos:

- ser datos reales, procedentes de mediciones, estimaciones, etc. sobre fenómenos reales
- tales fenómenos (y sus datos derivados) debían pertenecer a campos diversos del conocimiento. Como cubrir todo el espectro no es posible, al menos debían ser datos y fenómenos fáciles de entender y de interpretar por personas con procedencia e intereses académicos muy diversos
- como la utilidad de muchas técnicas SNA se evidencia cuando la cantidad de datos a tratar es elevada (las muestras pequeñas, en datos, pueden ser analizadas fácilmente de forma manual), las colecciones de datos deberían tener las dimensiones necesarias. Esto hace que, en muchos casos,

tales colecciones no puedan ser elaboradas de forma manual, sino a través de procedimientos automáticos; si bien en muchas ocasiones es preciso someter los resultados de esos procedimientos automáticos a controles de calidad, refinado, etc. de forma manual.

2.- Colecciones de datos

Se relacionan a continueción las colecciones construídas, algunas de sus características más señaladas y también las notas más señaladas sobre su proceso de construcción.

2.1 Universidades Europeas

Esta colección está construída a partir de los dominios web de las 100 universidades europeas mejor situadas en el *Web Ranking* de Instituciones Académicas Superiores (datos del primer semestre de 2013).

Para su elaboración se utilizó un *crawler* que de forma automática recorrió los dominios web de dichas universidades, recopilando páginas web y los enlaces que éstas contenían, especialmente los que apuntaban a páginas web de alguna de las otras 100 universidades consideradas.

La colección formada a partir de estos datos de base consiste, pues, en una red de 100 nodos (cada una de las universidades) con tres atributos: una etiqueta identificativa de cada universidad, el país de cada una y el número de páginas web de cada una de ellas, utilizable como indicador del tamaño de las mismas.

Los arcos de esa red son dirigidos y tienen un atributo (además de origen y destino, obviamente), el peso, considerado como la suma de los enlaces entre todas las páginas de cada par de nodos.

Estos datos se encuentran disponibles en formato *GraphML* (uno de los estándares en este campo). La red así formada es especialmente útil para análisis de:

- coeficientes de centralidad, comparación entre los coeficientes más usuales
- detección de comunidades (modularidad)

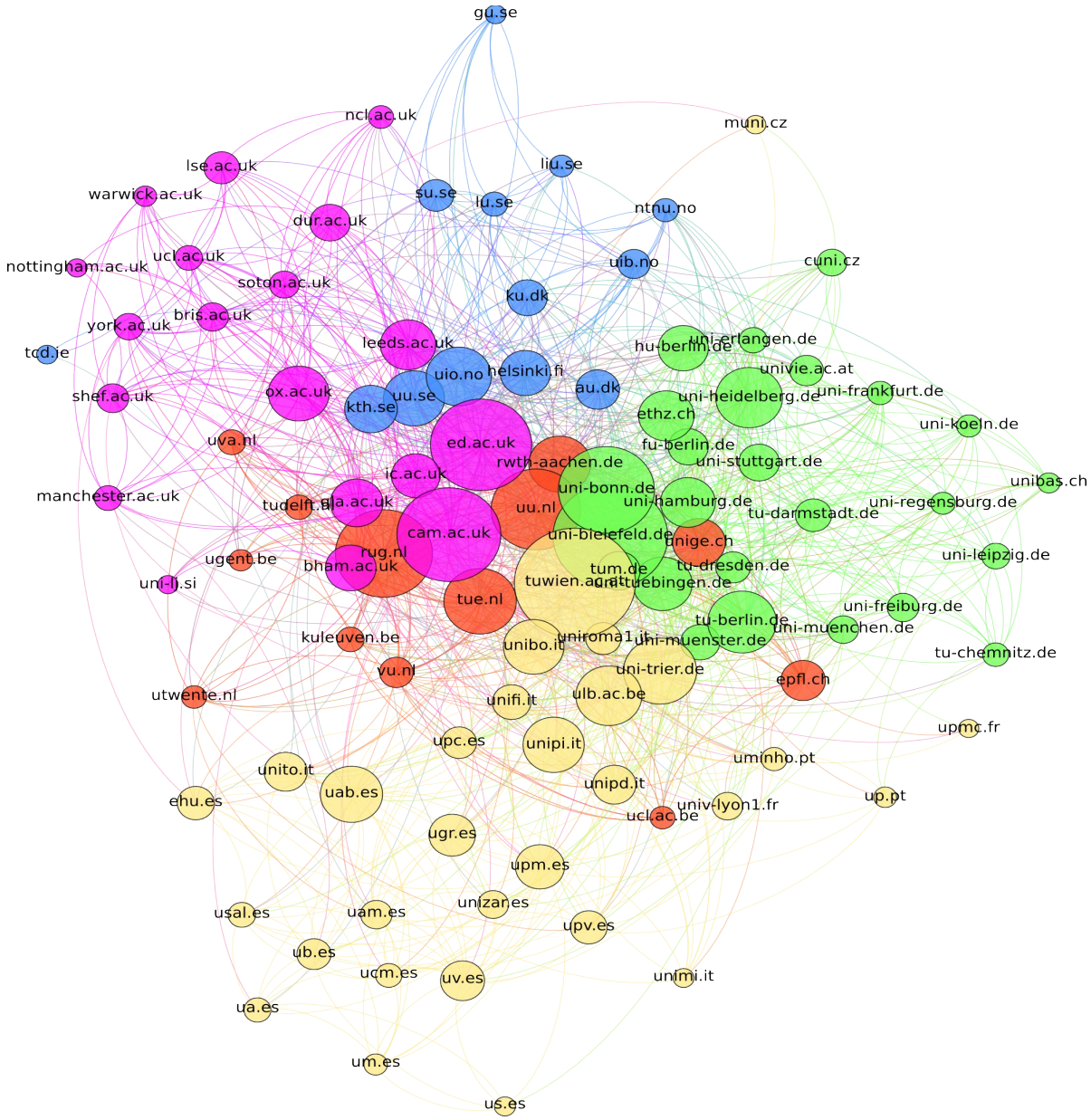

Figura 1: Universidades europeas, centralidad y tamaño

```
*Vertices 100
1 "uab.es" 339.41373 295.46432 0.0
2 "unige.ch" 239.65576 314.5294 0.0
3 "unipd.it" 269.3515 313.30246 0.0
4 "au.dk" 142.41043 326.84686 0.0
5 "uni-lj.si" 371.73697 91.64015 0.0
6 "shef.ac.uk" 382.4489 146.26082 0.0
7 "ox.ac.uk" 259.88947 204.67876 0.0
8 "unipi.it" 299.03006 346.98703 0.0
9 "unizar.es" 218.88974 427.17664 0.0
10 "uni-trier.de" 189.20297 355.1973 0.0
11 "tu-darmstadt.de" 89.7973 180.66107 0.0
12 "uni-regensburg.de" 33.97643 228.66068 0.0
13 "ub.es" 377.51254 388.4817 0.0
14 "ua.es" 361.60867 459.2596 0.0
15 "upv.es" 240.795 446.62973 0.0
16 "ucm.es" 315.91348 418.9094 0.0
17 "tu-berlin.de" 177.39201 290.31253 0.0
```

Figura 2: Datos en formato Pajek

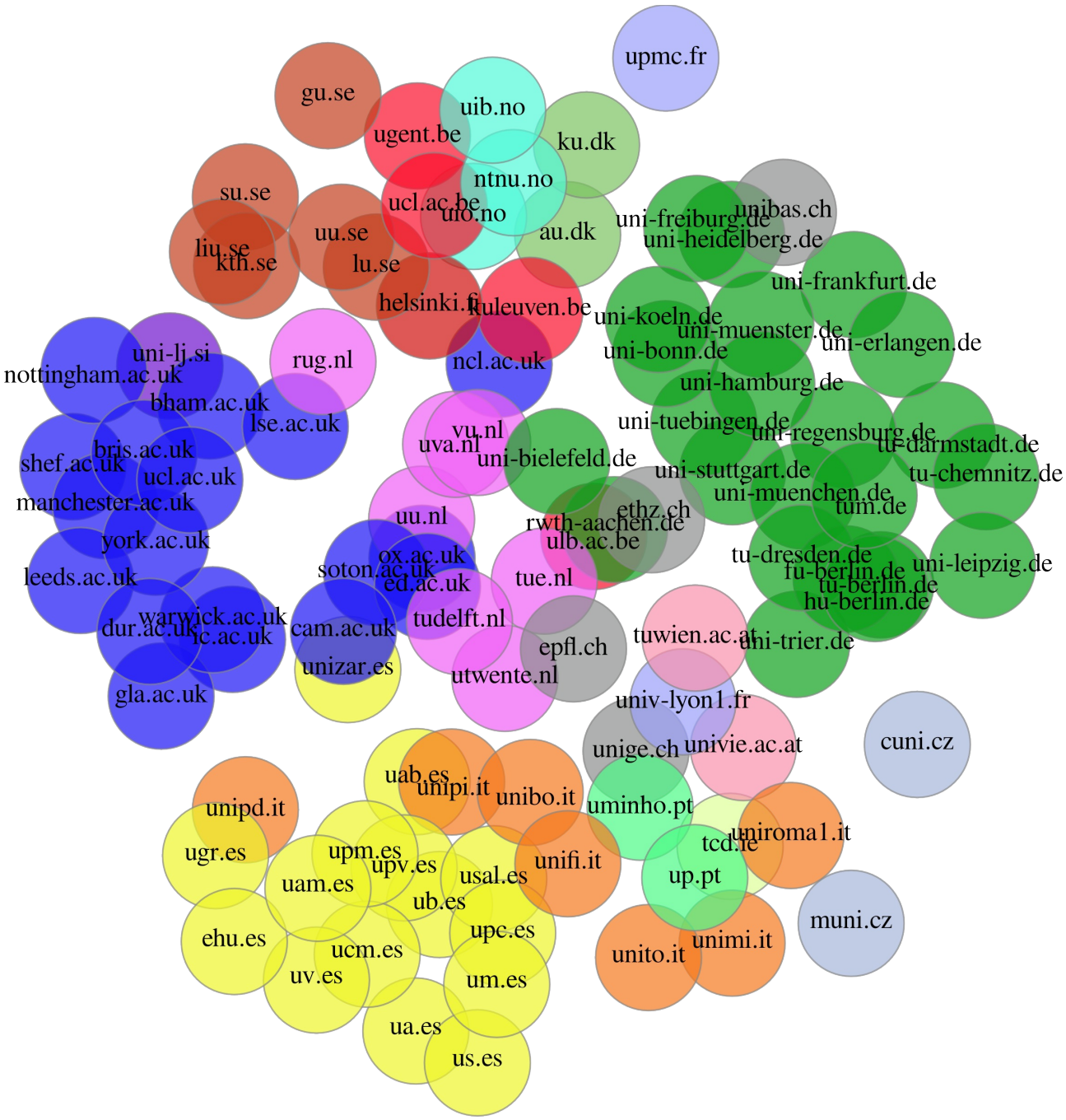

Figura 3: Universidades europeas, comunidades

2.2 Sitios web de Bibliotecas Nacionales de diversos países

Para formar esta colección se utilizó nuevamente un *crawler* que recopiló páginas y enlaces (en esta ocasión sólo internos) de diferentes bibliotecas nacionales de otros tantos países europeos. Para los objetivos propuestos, y teniendo en cuenta el tamaño, en general pequeño, de tales sitios web, se estimó que con las 1.000 primeras páginas de cada uno era suficiente.

Se formó así una red para cada una de las bibliotecas seleccionadas, representando directamente cada página como un nodo y los hiperlinks como arcos dirigidos.

Los datos están disponibles en formato *GraphML* y *Pajek*. Estas redes sirven para varios propósitos:

- dado su tamaño moderado, y que mapean directamente sitios web reales, es posible su navegación manual y la identificación de los nodos o páginas con mayores coeficientes de visibilidad, intermediación, etc.
- es posible la comparación de unas redes (bibliotecas) con las otras, permitiendo aplicar análisis topológico y su correlación con el tipo de institución y su estructura interna.
- la representación gráfica de tales redes permite analizar las diferencias entre unos algoritmos de representación y otros, así como el efecto del ajuste de diversos parámetros de dichos algoritmos

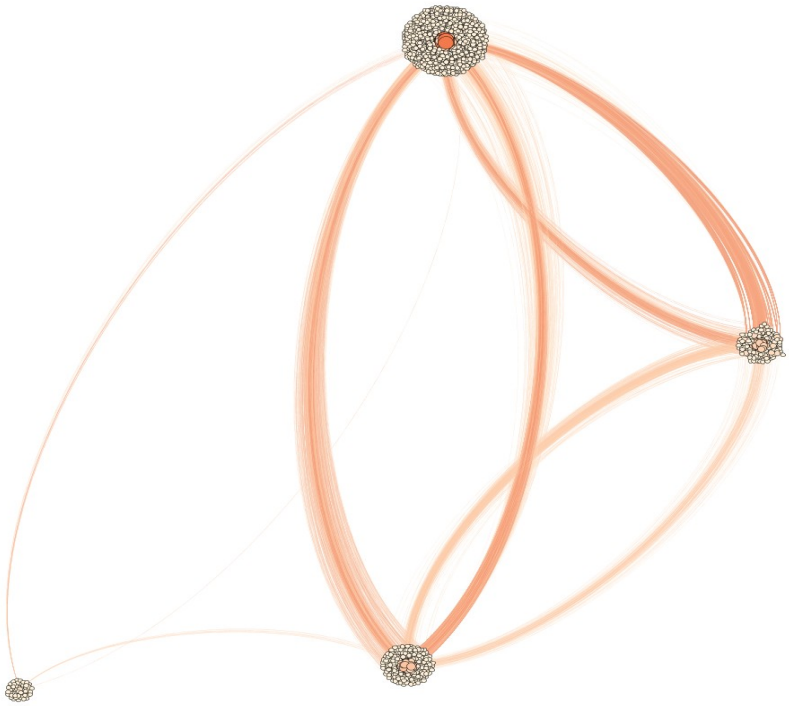

Figura 4: Grafo del website de la Biblioteca nacional de Francia

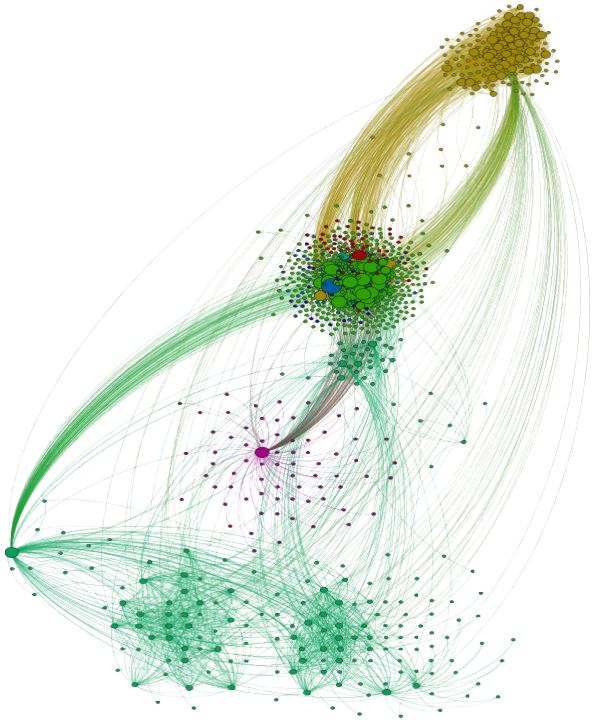

Figura 5: Website de la BN de Eslovenia

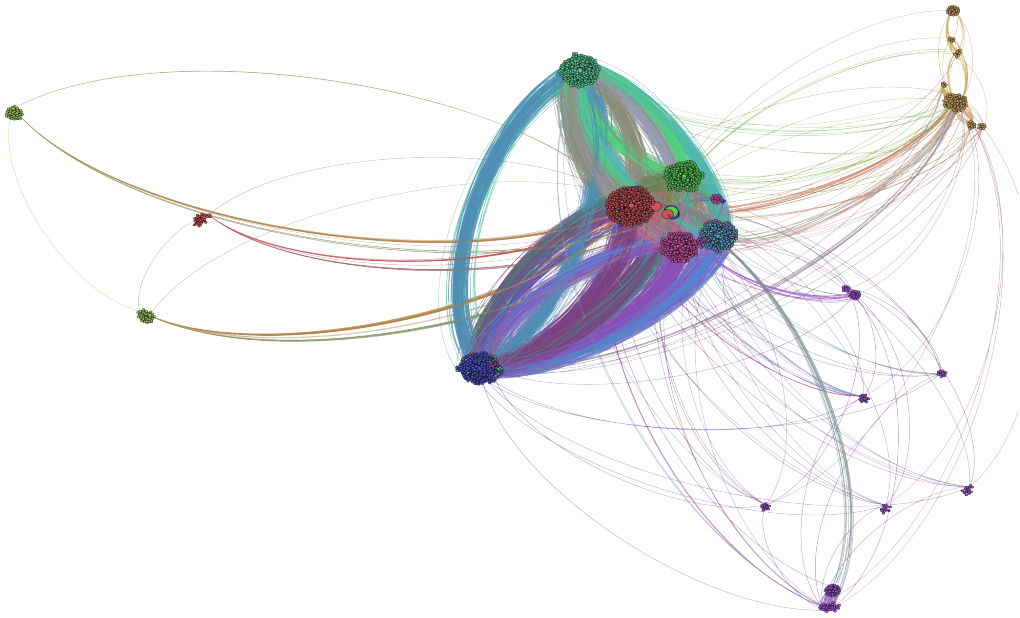

Figura 6: Website de la Biblioteca Nacional de España

2.3 Personas públicas de la Transición Española

Se trata de las personas que aparecen en todas las noticias de prensa de un importante periódico español (*El País*, 1977-1987), de todos los ámbitos. Las noticias fueron extraídas de forma exhaustiva de la hemeroteca de dicho diario y procesadas después un un reconocedor automático de entidades. Aunque este tipo de software no es preciso al 100 % y produce algunos errores, la elevada cantidad de datos utilizados minimiza la posible tasa de error.

En este caso, se consideró que dos personas que aparecían juntas en una misma noticia, del tipo y contenido que fuese, tenía una relación del agún tipo más fuerte en función de la frecuencia de tales coapariciones.

La red resultante (personas como nodos y coapariciones como enlaces no dirigidos y con peso) es de un tamaño realmente importante, y una buena muestra de la potencia de las técnicas SNA.

La red está en formato *CSV* (hubo que diseñar algunos *scripts* para su procesamiento e importación por parte de solftware especializado).

Esta colección de datos es especialmente interesante para:

- analizar los efectos de la aplicación de umbrales de pesado de los arcos
- estudiar métodos de detección de comunidades en redes grandes
- estudiar la utilidad y el significado de determinados coeficientes y medidas , especialmente los relacionados con la visibilidad, centralidad e intermediación
- analizar el efecto longitudinal y el manejo de la variable tiempo, dado que la red evoluciona al tener las coapariciones fecha.

```
Abigail_Folger Mia_Farrow
Abigail_Folger Rosemary
Adela_Barnés Carmen_Bravo-Villasante
Adela_Cortés Ana_María_Fanlo
Adela_Cortés Carmen_Herrero
Adela_Cortés Carmen_Llorca
Adela_Cortés Carmen_Martínez_Burquera
Adela_Cortés Diana
Adela_Cortés Francisca_Sauquillo
Adela_Cortés Josefina_Lobo
Adela_Cortés Lina_Ortas
Adela_Cortés Mari_Enma
Adela_Cortés Marta_Pastor
Adela_Cortés María_Dolores_Xoubanova
Adela_Cortés Paloma_González
Adelina_Vázquez Ana_María_Viéitez
Adelina_Vázquez María_Luisa_Areses
```

Figura 7: Datos en formato CSV (personas 1977-87)

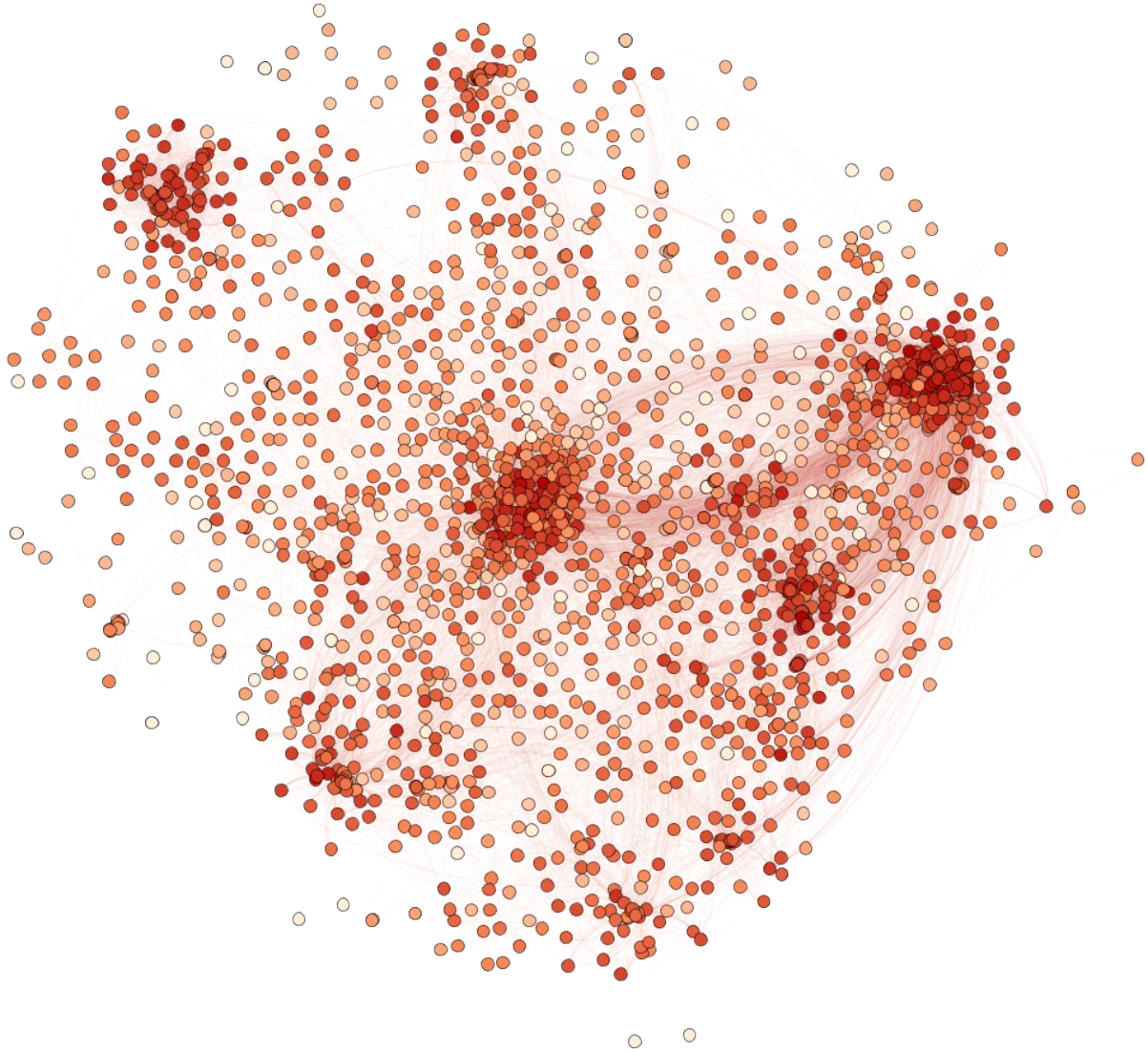

Figura 8: Grafo de personas 1977-87

3.- Utilidades de manipulación de datos

Para poder llevar a cabo la construcción de estas colecciones de datos hubieron de programarse diversas utilidades o herramientas, o bien, en algunos casos, readaptar otras existentes.

Así, se reforzó y se generalizó el crawler, que con anterioridad se había aplicado a otros usos, facilitando su manejo y configuración y, sobre todo, proporcionando una salida en formato *CSV*, más acorde con otros programas utilizados con frecuencia en SNA.

Se mejoró y completó un *script* de conversión de formatos, tomando como base o eje el formato *CSV*. Dicho *script* permite el paso o conversión a otros formatos frecuentes, en especial *GraphML* y

Pajek.

Igualmente, se implementaron diversas utilidades de cálculo de coeficientes, normalización, etc. en *Excel*, al ser éste un software ampliamente utilizado y conocido por personas de muy diversos orígenes académicos. También se ajustaron procedimientos de paso de datos de *Excel* a software específico SNA, y viceversa.

4.- Material docente complementario

De manera complementaria, se diseñaron varias presentaciones sobre las mencionadas técnicas SNA; en unos casos se trata de presentaciones nuevas y en otros se trata de la ampliación de presentaciones ya existentes con nuevos diagramas, ejemplos con las colecciones ahora diseñadas y ejercicios.

A modo de ejemplo, puede consultarse la presentación '*A walk on Python-igraph*' (http://grulla.usal.es/igraph), de acceso abierto. Otras presentaciones se encuentran en la plataforma Studium.

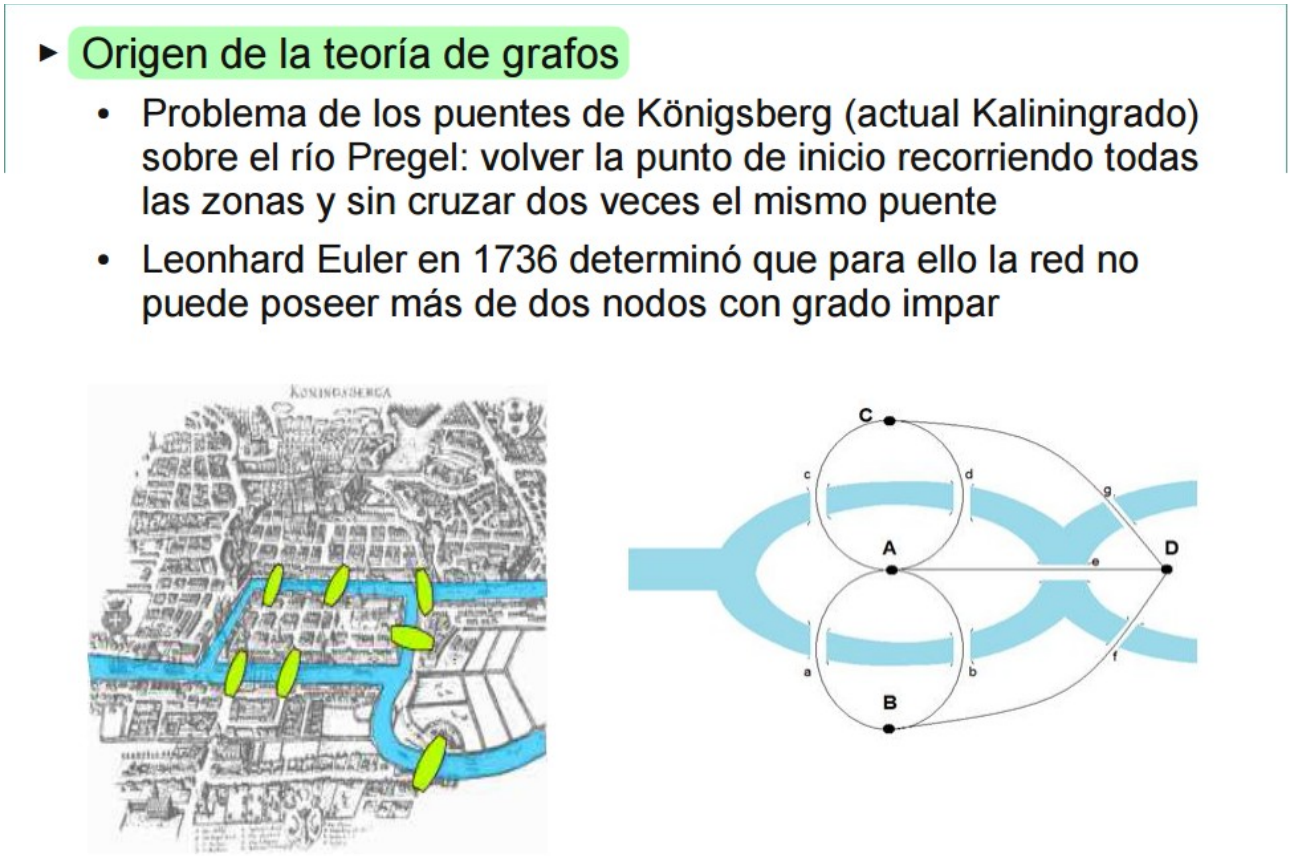

Figura 9: Slide de una de las presentaciones

- Grado de cercanía (*closeness*):
  - Mide la distancia de cada nodo con el resto de nodos
    - Se calcula dividiendo el número de nodos por el sumatorio de distancias entre el nodo y el resto de nodos
    - Los resultados más altos sugieren mayor facilidad de acceso al resto de nodos, con mayor capacidad para obtener y enviar información, y de manera más rápida

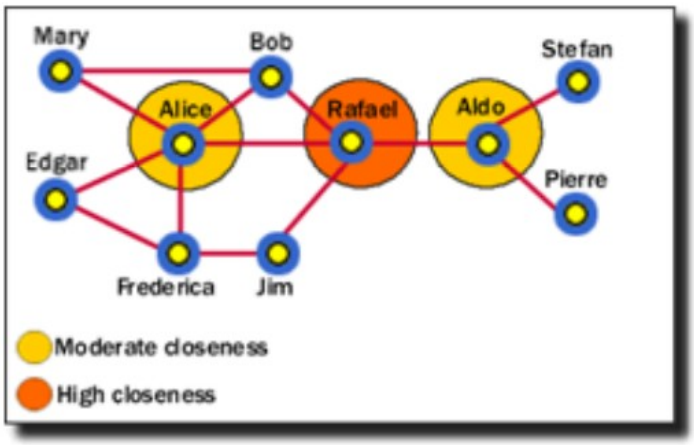

Figura 10: Ejemplo de una de las presentaciones